



Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers

Michael E. Mavroforakis^{a,*}, Harris V. Georgiou^a, Nikos Dimitropoulos^b,
Dionisis Cavouras^c, Sergios Theodoridis^a

^a University of Athens, Informatics Department, TYPA buildings, University Campus, 15771 Athens, Greece

^b Medical Imaging Department, EUROMEDICA Medical Center, 2 Mesogeion ave, Athens, Greece

^c Medical Instruments Technology Department, Technological Educational Institution of Athens, Egaleo 12210, Greece

Received 10 October 2005; received in revised form 23 March 2006; accepted 23 March 2006

KEYWORDS

Texture analysis;
Fractal dimension;
Mammography;
Medical diagnostics

Summary

Objective: Localized texture analysis of breast tissue on mammograms is an issue of major importance in mass characterization. However, in contrast to other mammographic diagnostic approaches, it has not been investigated in depth, due to its inherent difficulty and fuzziness. This work aims to the establishment of a quantitative approach of mammographic masses texture classification, based on advanced classifier architectures and supported by fractal analysis of the dataset of the extracted textural features. Additionally, a comparison of the information content of the proposed feature set with that of the qualitative characteristics used in clinical practice by expert radiologists is presented.

Methods and material: An extensive set of textural feature functions was applied to a set of 130 digitized mammograms, in multiple configurations and scales, constructing compact datasets of textural “signatures” for benign and malignant cases of tumors. These quantitative textural datasets were subsequently studied against a set of a thorough and compact list of qualitative texture descriptions of breast mass tissue, normally considered under a typical clinical assessment, in order to investigate the discriminating value and the statistical correlation between the two sets. Fractal analysis was employed to compare the information content and dimensionality of the textural features datasets with the qualitative information provided through medical

* Corresponding author at: 43 Knossou Street, P.O. 16561, Glyfada, Athens, Greece. Tel.: +30 210 9648663; fax: +30 210 9313631.
E-mail address: mmavrof@di.uoa.gr (M.E. Mavroforakis).

diagnosis. A wide range of linear and non-linear classification architectures was employed, including linear discriminant analysis (LDA), least-squares minimum distance (LSMD), K-nearest-neighbors (K-nn), radial basis function (RBF) and multi-layer perceptron (MLP) artificial neural network (ANN), as well as support vector machine (SVM) classifiers. The classification process was used as the means to evaluate the inherent quality and informational content of each of the datasets, as well as the objective performance of each of the classifiers themselves in real classification of mammographic breast tumors against verified diagnosis.

Results: Textural features extracted at larger scales and sampling box sizes proved to be more content-rich than their equivalents at smaller scales and sizes. Fractal analysis on the dimensionality of the textural datasets verified that reduced subsets of optimal feature combinations can describe the original feature space adequately for classification purposes and at least the same detail and quality as the list of qualitative texture descriptions provided by a human expert. Non-linear classifiers, especially SVMs, have been proven superior to any linear equivalent. Breast mass classification of mammograms, based only on textural features, achieved an optimal score of 83.9%, through SVM classifiers.

© 2006 Elsevier B.V. All rights reserved.

1. Introduction

The examination and analysis of mammographic images is a complex cognitive task that includes various aspects of medical expertise and conclusive clinical findings. The visual task of clinical evaluation and diagnosis, based on mammographic image screening, consists of a number of different factors in multiple scales and levels of decomposition, as approximately 80–85% of diagnostic information is retrieved from the appearance of the tumor itself [1–3]. The fine-scale organization of the informational content on the mammographic image is a key factor in the detection of malignancy, as it represents the nature and structure and, hence, the quality of biological tissues, as they are projected on the mammogram [4–6]. Similar textural features are also present in rare clinical cases, where direct inference on probably benignancy or malignancy is much more complex [7,8]. These fine-scale structural details are realized as visual patterns in the image and they are often referred to as “texture” of the corresponding region. When the mammogram is digitized with adequate quality and resolution, these textural patterns can be identified, analyzed and classified by a computer, based on the statistical properties and the spatial correlations between the pixel values [9].

In clinical practice, an experienced physician often identifies textural information in the form of qualitative characteristics and pathological findings, retrieved directly by examining the properties of the mammogram and combines them effectively with other data available from sources other than the mammographic image itself. In the context of successful clinical estimation, patient’s age and medical history have been proven issues of utmost

importance [10,11]. Furthermore, the presence of suspicious areas in the form of tumors is often examined by investigating the textural content of the mammographic image [12,13]. Another significant property is the presence and morphology of microcalcifications, as well as the shape and the morphology of the tumor itself [14–18].

Many studies have been focused on the general issue of texture analysis on mammographic images, in the context of detection of the boundary of tumors and microcalcifications [19,20], since the task of localization of tumors or microcalcifications is, apart from attractive – due to its apparent usefulness – comparatively easy to implement, as it relies on the notable difference in contrast and intensity of the target from its surrounding parenchyma and therefore, simpler processing and classification methods are needed. However, in the task of textural mass characterization, due to its inherent difficulty, the selection of optimal parameters and configuration for the various textural feature functions is still model-restrictive and case-specific [21–24]. Furthermore, there are only a few studies available, related to the more general issue of investigating the inherent complexity of the extracted textural data and the suggestive structure of the corresponding training datasets for classifiers [25].

This study is aiming at filling this void and is focused on three main areas of interest: (a) the investigation of various properties of common textural features in relation to the sampling size and scale, (b) the investigation of the inherent complexity of texture datasets, using statistical and fractal analysis techniques, and (c) the application of LDA, LSMD, K-nn, ANN and SVM classifiers for the evaluation of their efficiency in real diagnostic applications.

For the purposes of this study, an original mammogram database was studied in the context of verified clinical results. The database contained detailed qualitative information for each mammogram, including a thorough and compact list of clinical findings provided by a human expert. The statistical significance of the diagnostic discriminative information for each one of these components, both separately and in combination, has already been established [26]. An extensive set of textural feature extractors, in various configurations and scales, has been applied on the mammogram database in order to construct complete datasets of textural “signatures” for benign and malignant cases of tumors. Multivariate analysis of variance (MANOVA) [27,28] was also applied for the construction of additional subsets of statistically independent textural features. Both the original and the reduced datasets were analyzed using statistical analysis and fractal dimension techniques [25], in order to establish a lower bound for the inherent dimensionality of the input space and how it is affected when using feature selection methods like MANOVA. Finally, the texture datasets for various sampling box sizes and scales were applied in a wide range of linear and non-linear classifier models, in order to evaluate the objective performance of each of the classifiers, as well as the inherent quality and informational content of each of the datasets in real classification problems.

The core material presented in this study is organized in sections as follows. Section 2 contains all the details regarding the base dataset and the methodologies used throughout the study. First, the digitized mammographic database is described in detail. The complete method of the textural analysis and the corresponding parameters for the localized image processing (e.g. sampling box sizes) is established in Section 2.3, while Section 2.4 describes the list of functions used as textural feature extractors. Section 2.5 illustrates a fractal-based method for estimating the dimensionality and the complexity of the feature space of a dataset in a quantitative way, such that it can be used as a method for estimating the descriptive power and content-richness of individual or combined textural features. Finally, Section 2.6 lists the range of linear, neural and SVM classifiers used in the final part of the study.

Next, Section 3 presents all the results of the textural features evaluation, the dataset fractal analysis and the classification tests. Section 3.1 consists of two parts, namely (a) the significance analysis of the feature selections with respect to their descriptive power and information content, and (b) the issues related to the significance of

sampling box sizes and feature extraction parameters. The quality and optimality of feature selections are both investigated using statistical significance analysis, as well as real classification runs. Section 3.2 illustrates the detailed results of the fractal analysis of datasets; their descriptive power is analyzed thoroughly in both their original and reduced-dimensionality versions, thus validating the use of optimal subsets of the textural features instead of all of them. Additionally, fractal analysis was performed on the dataset of the qualitative features provided by an expert physician on the same mammograms, in order to compare the intrinsic information content and dimensionality of the proposed feature set and the qualitative features used in practice by the experienced radiologists, during the diagnostic process. Furthermore, classification results and comparative classifier performance is presented in Section 3.3, thus proving the practical value and descriptive power of the selected textural features for diagnostic purposes.

Finally, Section 4 discusses the choices and implications of various aspects of the extraction of textural features, the fractal-based analysis of the datasets and the performance of the various classifiers, while Section 5 summarizes the consequences of using similar methods for automated image analysis and computer-aided diagnosis. The study is enriched with two appendices, one with a complete list of mathematical formulations for all the statistical functions used as textural feature extractors and one with a brief description of the Tukey windowing function, used in fractal analysis.

2. Material and methods

According to the scope of this study, i.e., the textural features analysis, the dataset fractal analysis and the classification models, six distinct resource materials were used: (a) a prototype mammographic image database, (b) the corresponding dataset of qualitative features provided by the expert physician, (c) a thorough set of textural feature functions, (d) multi-level localized image processing, (e) textural features extraction, (f) dataset fractal dimension evaluation, and (g) linear, neural networks and SVM classification architectures.

2.1. Digitized mammographic database

The requirement for patients clinical history and positive histological verification of the benignancy or malignancy of each case was assessed as one of

extreme importance for the quality and validity of the subsequent results. Thus, a new special-purpose image set was assembled, using cases of mammographic tumors with complete radiological evaluation and histological diagnosis [26]. The initial set contained 1350 mammograms of women that proceeded for the evaluation of a clinical breast problem and it was used as a base for the final selection of tumor cases with positive clinical verification by surgical biopsy and histological examination. For the construction of the raw material for this study, a set of 130 mammograms were selected for digitization by an expert physician, containing a total of 46 benign cases and 84 tumor malignancies of various types. The selection was made on the basis of unbiased statistical distribution of the underlying textural features and the completeness of the dataset with respect to various clinical findings, always focused in pathological cases that included presence of at least one suspicious mass. In the set of 130 mammograms, almost two-thirds of the cases are malignant because many different types of malignancies had to be represented in the database with an adequate statistical sample. For each mammogram, a complete list of qualitative information was provided by the attendant physician [26], containing details about the age of the patient, presence and number of tumors, microcalcifications, density of the tumor, percentage of fat inside the mass, tumor boundary vagueness, tumor homogeneity, tumor shape type and clinical diagnosis. All these qualitative clinical descriptions, as well as their logging range of values, is presented in Table 1. The various qualitative details were included according to the expert's recommendations and proposals, as explicit information related to various types of malignant mammogram abnorm-

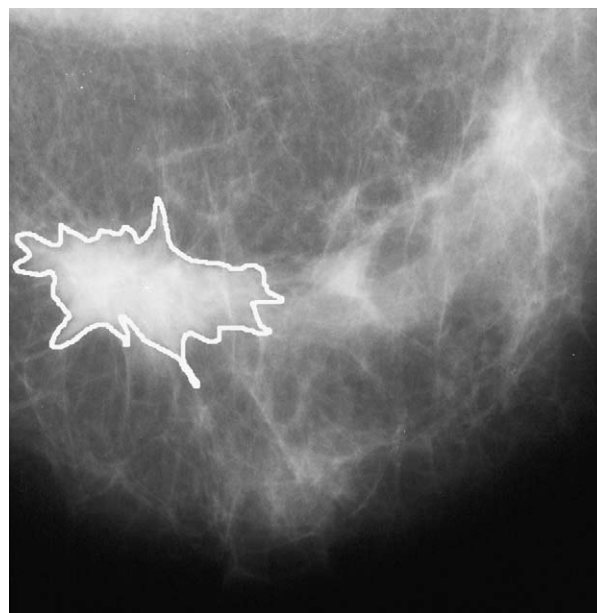


Figure 1 Sample digitized mammogram with one manually segmented suspicious mass.

alities, including architectural distortion, microcalcification clusters, micro-lobulated or stellate masses [10,11,29].

The mammograms were digitized at a resolution of 63 μm (400 dpi) at 8-bit gray level, which is consistent with other typical image databases of digitized mammograms that are used as a reference in similar studies [30]. The final set of 130 mammograms was used in all cases with no reduction in spatial resolution or gray level depth.

Advanced algorithms for automated mammographic lesion detection have been proposed, however their level of sensitivity and specificity is still under investigation [21–24]. Furthermore, their fine-scale accordance to the corresponding expert's detailed description of tumor boundaries, especially in cases of non-trivial stellate distortions, exhibits many practical problems [22]. In order to obtain tumor boundaries of high quality and detail, a manual segmentation was applied. Specifically, each tumor was manually described by the radiologists at the maximum available spatial resolution (Fig. 1), using a high-resolution digitizer device and stored as an embedded boundary descriptor via alpha channel data. These tumor descriptions were subsequently used as a fine-scale tissue inclusion/exclusion mask for all subsequent extraction and analysis of textural features.

2.2. Datasets and features

A large set of textural feature functions were applied in multiple configurations, in order to

Table 1 Complete list and quantification details for all the qualitative properties, used by the expert physician when annotating the images in the mammogram database

Qualitative feature	Range
Patient's age	Years (integer)
Mass existence	Yes/no
Microcalcifications existence	Yes/no
Fat percentage	0%, ..., 100%
Boundary sharpness	0%, ..., 100%
Mass density	L (hypo)/M (iso)/H (hyper)
Mass Homogeneity	1, ..., 10
Mass shape type	1 (round)/2 (lobulated)/3 (micro-lobulated)/4 (stellate)
Verified diagnosis	B (benign)/M (malignant)

investigate the relative effect of the quality of texture information extracted from the digitized images. The extracted feature values were normalized against average and standard deviation for each dimension of the dataset using Gaussian distribution approximation, according to the following formula:

$$y_i = \frac{x_i - \bar{X}}{\sigma} \quad (1)$$

where \bar{X} and σ are the mean value and the standard deviation N -dimensional vectors, respectively.

It should be noted that, although feature values within the datasets were normalized according to Eq. (1), the pixel values of the images were not. This means that the textural features were calculated upon the original grayscale pixel values, without any histogram or other modification. The reason for this is the fact that, with the exception of the 6 first-order statistics, all the other textural feature functions presented in Appendix B are resilient to exact pixel values. Instead, they calculate statistical properties upon the co-occurrence and run-length matrices of the pixel values, i.e., they highlight correlations and relative differences between the values rather than upon the absolute values themselves. Even if the original images refer to different X-ray exposure parameters, the content and physical meaning of these textural features relates to the inherent micro-level structural details of the underlying tissue, not the absolute values of individual pixels.

For the six functions of the first-order statistics, exposure level is indeed important for the overall brightness of the final image. However, since all mammograms were acquired using optimal X-ray exposure settings for the entire breast area, the gray-level histogram profile within the tumor region should be more relevant to the actual properties of the underlying tissue itself (e.g., density, homogeneity, etc.), rather than any explicit differences in the exposure rates. Furthermore, the same first-order statistics for the entire breast area are also included in the datasets, so that these intra- and extra-tumor statistics can be used in conjunction or comparatively by a classifier.

In all cases, normalizing each feature (i.e., dimension) by Eq. (1) ensures that the datasets used for training and testing the classifiers are guaranteed to exhibit zero-mean normal distributions, for best results in neural and SVM classifiers [31,32].

2.3. Localized image processing

Each mammographic image was first segmented into mass and non-mass regions according to the tumor's

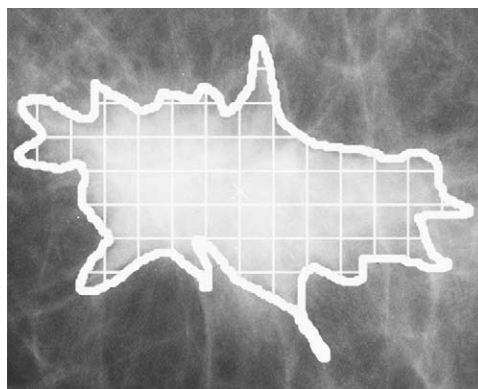


Figure 2 Suspicious region of interest and relative size of 50-pixel box for texture analysis, no box-interleaving.

boundary description provided by the experienced physician. The identified mass tissue areas were further divided according to two configurations, into continuous sub-regions of sampling boxes and sizes of 20 and 50 pixels, respectively. The specific sampling box sizes correspond to spatial resolution of 1.270 and 3.175 mm of mass tissue, respectively, which were asserted by the expert as segments of size adequate to capture significant textural information content in mammograms. Clinical studies have established that the expected diameter of tumors ranges from 3 up to 30 mm [21], or roughly 48–476 pixels for the specified image scanning resolution (63 μm). This means that sampling box sizes larger than 3 mm or roughly 50 pixels wide would be too large for some tumors. Additionally, when statistical features are calculated over increasingly large areas of a digital image, the results refer more to the macro-scale morphological and structural properties rather than the micro-scale textural properties of the image content [9]. Similarly, any box size significantly smaller than 3 mm or 50 pixels should also be discarded as it would not refer to any valid tumor but to other features at smaller scales, such as microcalcifications or noise. Therefore, a typical size of 50 was decided as a standard baseline for the sampling box (Fig. 2), plus one much smaller size at 20 pixels for investigating the differences in quality and content of the textural features when the size of the sampling box changes.

In the configuration of sampling box size of 1.270 mm (20 pixels), the sampling region was significantly lower than the minimum expected tumor size of 3 mm. Consequently, exhaustive texture sampling was considered redundant and a box-interleaving scheme of one-by-one in each dimension was employed (Fig. 3). The discriminating value and statistical significance of the textural features were not affected by this sub-sampling scheme, since the

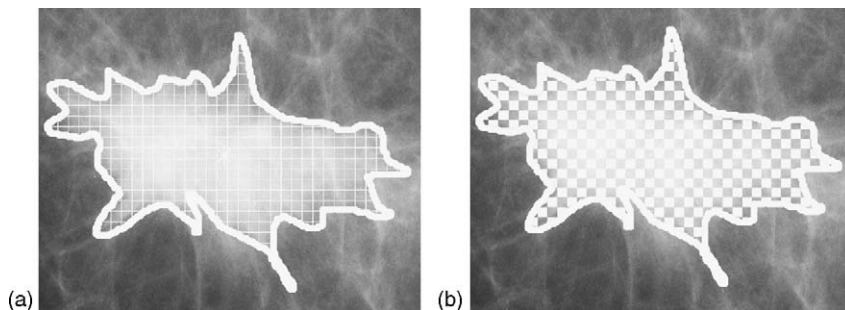


Figure 3 (a) Suspicious region of interest and relative size of 20-pixel box for texture analysis with box-interleaving. In this case, only every second box is used for textural features extraction, i.e., only the non-white areas in (b).

size of the smallest object of interest, i.e., a tumor of minimum diameter, is still more than twice as big as the size of the 20-pixel sampling box. In the case of 3.175 mm sampling box size (50 pixels), i.e., roughly the same size as the smallest expected tumor diameter, no box-interleaving scheme was applied in order to preserve the complete informational content of each image sample.

For each sampled box, grayscale co-occurrence [33,34] and run-length [35] matrices were computed for three distinct neighboring pixel configurations, according to a distance factor $d \in \{1,2,3\}$. The application of each one of the three pixel-neighboring configurations throughout the entire spatial matrix of the current sampled box created three corresponding sampling modes at pixel-level, essentially affecting the exact pixels upon which the textural feature functions were applied (Fig. 4). The reason for using multiple distance factors was to evaluate the effect of sampling at various pixel levels, in relation to the quality and consistency of the extracted textural features, as well as the investigation of different information content, captured at different scales between neighboring pixels.

For each different pixel-neighboring configuration, all the available (primary) directions were used for the calculation of co-occurrence and run-

length matrices. The number of available angular directions is equal to the half of the total number of pixels found at distance d from the point under examination (i.e., the center pixel) of the current box (Fig. 4). That is because any two opposing peripheral pixels define a unique angular direction. Consequently, for $d = 1$ there were 4, for $d = 2$ there were 8 and for $d = 3$ there were 12 (unique) angular directions available, i.e., over the numbered cells in graphs (a), (b) and (c), respectively, in Fig. 4. The direct effect of having more than four angular directions (two orthogonal and two diagonal) is essentially the computation of the co-occurrence and run-length matrices over more directions, i.e., more samples. The corresponding textural features are now calculated over more image data, without altering their statistical, physical or image-related properties. Thus, this scheme enhances the informational content and the quality of subsequent statistical properties of these features, using increased scale and tissue sampling area.

For each angular direction, the complete set of the available texture functions was computed. The average and range (min–max) of the feature values over all the available directions were stored for each specific distance factor ($d = 1,2,3$) as a compound textural “signature” of every image sample.

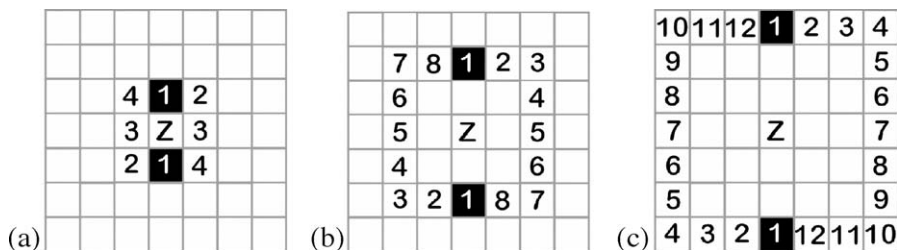


Figure 4 Calculation of texture statistics and gray-level correlations over various pixel-neighboring distances. Cells indicate pixels, the point Z indicates the current pixel within the sampling box and numbers enumerate the angular directions in each case, moving in clockwise direction. Using point Z as the center, each pair of cells with the same number indicate a unique angular direction. Every such pair of pixels is then used to update the gray-level spatial co-occurrence matrix. For first-order and run-length statistics, the pixel value at point Z is used instead.

The result was the creation of two compound texture datasets, namely for sampling boxes of size 1.270 mm (20 pixels) and 3.175 mm (50 pixels), each containing a complete set of textural features over the three pixel-neighboring configurations, as well as a set of typical first-order gray-level statistics for each complete image. Appendix A contains the list and the mathematical formulation of all first-order, second-order and run-length statistics that were used as textural feature functions for the creation of these datasets.

2.4. Textural features extraction

From each sampled sub-region, texture information was examined by extracting first-order statistics, second-order statistics and gray-level runs. All subsequent analysis preserved the original gray-level and spatial resolution, and all runs were examined up to their full length with no concatenation or limits in maximum runs.

First-order statistics of the gray-level distribution for each image sub-region matrix $I(x,y)$ were examined through six commonly used metrics proposed by Haralick et al. [33,34,36]. Namely, min value, max value, average value, variance, skewness and kurtosis were used as descriptive measurements of the overall gray-level histogram of the mass. The complete list and analytical formulas of these statistics are presented in detail in Appendix A.

Second-order statistics of the gray-level distribution, derived from spatial distribution gray-level matrices (SDGM), were examined through 14 commonly used metrics also proposed by Haralick et al. [33,34,36]. The complete list and analytical formulas of all the second-order statistics are also presented in detail in Appendix A.

Gray-level runs, derived from run-length matrices (RLM), were examined with the introduction of five commonly used run-length metrics, proposed by Galloway [35,36]. Namely, short runs emphasis, long runs emphasis, run-length non-uniformity and run percentage were used as descriptive measurements of each of the run-length matrix calculated over the sampled image sub-regions. The complete list and analytical formulas of all the run-length statistics are also presented in detail in Appendix A.

From the 14 second-order statistics features and the five run-length descriptive measurements, only the average and the range of values over all the distinct angular directions (per each of the three neighboring distances) were computed and stored in the texture datasets. Thus, 38 calculated values were utilized (per each of the three neighboring distance configurations encountered) as features

representing the second and higher order statistics of mammograms texture. Combined with the 6 first-order statistics feature measurements, the total number of features extracted for each sampled sub-region of the image was 120.

A separate dataset was created for every sampling box size, namely 1.270 mm (20 pixels) and 3.175 mm (50 pixels). These datasets were subsequently used for fractal dimension analysis of the texture, as well as by the classifiers for both testing and training purposes, employing standard k -fold cross-validation techniques for partitioning the complete sets into distinct subsets [31,37–39].

2.5. Fractal dimension analysis

In order to establish a preliminary estimation of the complexity and intrinsic dimensionality of the texture datasets, fractal feature analysis was applied. Fractal feature analysis, specifically the calculation of the *intrinsic fractal dimension* of the input datasets, provides the quantitative means of investigating the degree of linear independence and the correlation between the available features by means of dimensionality of the resulting feature space [40,41]. Fractal dimension has also been used as an alternative way of characterizing the discriminative power of feature combinations, thus providing a non-statistical way of ranking them in terms of importance for the classification task [25]. The two most commonly used methods of calculating the fractal dimension of a dataset are the *pair-count* (PC) and the *box-counting* (BC) algorithms [41–43].

In *pair-count* algorithm, all Euclidean distances between the samples of the dataset are calculated and a closure measure is then used to cluster the resulting distances space into groups, according to various ranges (r), i.e., the maximum allowable distance within samples of the same group. The $PC(r)$ value is calculated for various sizes of r and it has been proven that $PC(r)$ can be approximated by

$$PC(r) = K \cdot r^D \quad (2)$$

where K is a constant and D is called the pair-count exponent. The $PC(r)$ plot is then a plot of: $\log PC(r)$ versus $\log(r)$, i.e., D is the slope of the linear part of the $PC(r)$ plot over a specific range of distances (r). The exponent D is called *correlation fractal dimension* of the dataset, or D_2 .

The *box-counting* approach is commonly used when the datasets contain large number of samples, usually in the order of thousands [25,41]. In this case, instead of calculating all distances between

the samples, the input space is partitioned into a grid of n -dimensional cells of side equal to r . Then, the samples inside each cell are calculated and the *frequency of occurrence* (C_r), i.e., the count of samples in a cell, divided by the total number of samples, is used to approximate the correlation fractal dimension by

$$D_2 = \frac{\partial \log \sum_{(i)} (C_r^i)^2}{\partial \log(r)} \quad (3)$$

Ideally, both pair-count algorithm and box-counting algorithm calculate the same value, i.e., the correlation fractal dimension D_2 of the initial dataset, which characterizes the intrinsic or “true” dimension of the input space [41]. In other words, D_2 would be the *minimum dimension of the dataset* if only “perfect” features were allowed, i.e., totally uncorrelated and with the best discriminative power available within the specific set of features.

In this study, fractal analysis was applied to both the initial set of qualitative characteristics, provided by the expert physician, as well as the constructed datasets of textural features, in order to compare the information content of each set. In all cases, the pair-count algorithm employing Euclidean distances was used, due to the relatively small number of samples available, as well as the better stability and accuracy for D_2 against the box-counting approach.

For better accuracy, a parametric sigmoid function was used for fitting between the points of the PC(r) plot, in order to calculate the slope of the linear part. In the parametric sigmoid function

$$y = Y_0 + C_y \cdot \left[\frac{1}{1 + \exp(-C_x \cdot (x - X_0))} \right] \quad (4)$$

(X_0, Y_0) identifies the transposition of the axes, while C_x and C_y identify the appropriate scaling factors. Specifically, the value of C_x affects the steepness of the central part of the curve, while C_y specifies the Y -axis width of the sigmoid curve. Then, the slope of the linear part around the central curvature point, i.e., the value of D_2 , is

$$\frac{\partial^2 y(X_0)}{\partial x^2} = 0 \Rightarrow D_2 = \frac{\partial y(X_0)}{\partial x} = \frac{C_x \cdot C_y}{4} \quad (5)$$

The fitness of the parametric sigmoid over a range of samples assumes uniform error weighting over the entire range of data. Thus, if a large percentage of points lies near the upper bound ($y = Y_{\max}$) or lower bound ($y = Y_{\min}$) of the Y -axis range, as in most cases of PC(r) plots, then the fitness in the central region of the sigmoid, i.e., where the slope is calculated, can be fairly poor. For this reason, an additional

weighting factor was introduced in the fitness calculation in this study. Specifically, the *Tukey* (tapered cosine) parametric window function [44,45] was applied over the Y -axis range when calculating the overall fitness error of the sigmoid. The Tukey window is parametric (q) in terms of the exact form around its center, ranging from completely rectangular to completely triangular or *Hanning* window. When applied over the Y -axis range, the rectangular case ($q = 0$) is equivalent to calculating the fitness error uniformly over the entire range, while the triangular case ($q = 1$) is equivalent to calculating the fitness error primarily against the central point of the sigmoid curve. In this study, all fitness calculations employed Tukey windows as error weighting factors, using parameters q in the range between 0.5 and 1.0 for optimal slope results. The exact formula and details of the Tukey window is presented in [Appendix B](#).

2.6. Classification and testing

Although the textural features contained in the two initial datasets could be used for the estimation of any one of the qualitative data ([Table 1](#)) except patient’s age, classifications were conducted against clinical diagnosis only, as it is the dominant data component required in most clinical cases.

Several classifier architectures were applied during the classification phase. A LDA model was used in the form of linear classifier [46]. A LSMD was employed, using Mahalanobis distance measure and least-squares dataset pre-processing on the input [31,47]. A K-nn model was also used, including estimation of an optimal value K for the size of the neighborhood set [31].

Two different types of neural network architectures were employed: a RBF ANN with Gaussian activation functions and linear output functions [48], and a MLP ANN with hyperbolic tangent internal activations and softmax output functions [49], both implemented with topology adapted to each configuration and dataset. All topologies included one hidden layer of optimized size.

For more advanced investigation of the feature set, typical SVM models were applied in relation to the final diagnosis. Specifically, the C-Support Vector Classification (C-SVC) model was used in combination with standard RBF kernel functions, optimizing the penalty factor (C) and the Gaussian spread parameter (σ) during training [32]. SVM classifiers that employ various non-linear kernels are considered state-of-the-art in Pattern Recognition today [31,50] and they can be regarded as a realistic upper limit in the performance of automated systems in similar applications in practice.

For practical classification applications of high dimensionality, k -fold cross-validation techniques, specifically leave-one-out and leave- k -out methods [31,37–39], are usually employed in order to compensate with relatively small input datasets that have to be used both for training and testing. In this study, all configurations used leave-one-out method for dataset manipulation during training and testing phases, combined with optimal feature set selection for the linear classifiers, or the complete selected (optimal) feature sets for the neural networks. SVM classifiers employed limited feature set optimizations, using iterative runs of enlarging inclusions of several features, available on the feature ranking lists created by MANOVA significance analysis. One of the reasons that full feature set optimization was not applied with ANN or SVM architectures, is that the training phase, combined with the optimization of the size of the hidden layer in the case of ANN, becomes computationally too expensive. Furthermore, it also relies on the fact that trained ANN architectures apply optimal weight values at the input layer, thus they can be examined during a post-training pruning phase to optimally reduce the dimensionality of the input set if necessary [51]. SVM classifiers have also proven exceptionally efficient in classification problems of high dimensionality [31,32,47]. In all cases, classifiers were re-trained for every new dataset that was produced after the extraction of one training sample, according to the leave-one-out method, and then classified this sample treating it as unknown input. Thus, the quality and generalization of the classification results depended solely on the quality and unbiased distribution of the training samples in the complete dataset for each case.

3. Statistical analysis and classification results

The results of this study can be grouped into three main categories, one for each of the initial scopes of investigation: (a) the effect of different textural feature functions and configurations, in relation to sampling image sub-regions box sizes and pixel-neighboring distance values (scales), (b) the evaluation of the intrinsic descriptive power of each of the datasets via fractal dimensionality analysis, and (c) the performance of a wide range of linear, neural and SVM classifiers in real diagnostic applications.

3.1. Textural features and configurations

Within the scope of textural features analysis, the two initial texture datasets (box sizes of 20 and 50)

were employed with one or more pixel-neighboring distances. The study was focused in two issues: (i) the statistical significance analysis and feature selection using MANOVA, and (ii) the effects of using smaller (20-pixel) or larger (50-pixel) sampling boxes, as well as the effects of using one ($d = 1$) or more ($d = 1,2,3$) pixel-neighboring distances.

3.1.1. Statistical significance analysis and feature selections

Due to the high dimensionality of the initial texture datasets, each containing 120 discrete features, the MANOVA method was employed to select the most prominent features in a statistically independent way. Features from each texture dataset were rated and consequentially sorted by applying MANOVA significance analysis. From the resulting ranked sets, subsets of the best 10–20 features were evaluated in real classification cases against verified diagnosis, using all the linear and non-linear classifiers.

Results from the initial feature rankings were further investigated through exhaustive search for optimal combinations, as estimated against verified diagnosis by LDA and LSMD classifiers. As a result, the two lists provided by MANOVA were annotated according to these optimized feature subsets, underlying the features participating in one or more of these optimized feature subsets.

Tables 2 and 3 present the 10 best MANOVA selections for each of the initial texture datasets, i.e., for sampling box size of 20 and 50 pixels, respectively. Underlined features denote the ones that were selected within optimal feature subsets in both the LDA and the LSMD classification setups.

3.1.2. Sampling box sizes and pixel-neighboring modes

During the textural features analysis, six classification configurations were used in total. Specifically,

Table 2 Top-10 feature selections for the texture datasets of 20-pixel sampling box

Feature	Description
<u>86</u>	SDGM contrast, mean, $d = 3$
<u>4</u>	Gray-level mean value
<u>9</u>	SDGM, angular 2nd moment, range, $d = 1$
<u>80</u>	RLM non-uniformity, mean, $d = 2$
<u>82</u>	RLM percentage, mean, $d = 2$
<u>50</u>	SDGM correlation, mean, $d = 2$
<u>92</u>	SDGM invar. diff. moment, mean, $d = 3$
<u>74</u>	RLM short run emphasis, mean, $d = 2$
<u>73</u>	SDGM max. correl. coeff., range, $d = 2$
<u>120</u>	RLM percentage, mean, $d = 3$

Underlines indicate features also selected in optimal feature combinations by LSMD and LDA classifiers.

Table 3 Top-10 feature selections for the texture datasets of 50-pixel sampling box

Feature	Description
<u>80</u>	RLM non-uniformity, mean, $d = 2$
<u>20</u>	SDGM sum of variances, mean, $d = 1$
<u>3</u>	Gray-level max value
<u>90</u>	SDGM sum variance, mean, $d = 3$
<u>81</u>	RLM non-uniformity, range, $d = 2$
<u>6</u>	Gray-level histogram skewness
<u>15</u>	SDGM sum variance, mean, $d = 1$
<u>36</u>	RLM short run emphasis, mean, $d = 1$
<u>63</u>	SDGM entropy, range, $d = 2$
<u>72</u>	SDGM max. correl. coeff., mean, $d = 2$

Underlines indicate features also selected in optimal feature combinations by LSMD and LDA classifiers.

the classification tests for optimizing feature selection included the 20-pixel and the 50-pixel box size cases, at pixel distances $d = \{1\}$ and $d = \{1,2,3\}$. For better comparison between the different pixel-neighboring modes, any features on first-order statistics were permitted in a third separate case, essentially including all 10 of the initial MANOVA selections. Evaluation was conducted by using LDA and LSMD classifiers, exploiting all possible combinations of the current feature set and applying leave-one-out testing for evaluating the performance of each classifier on the complete texture datasets.

For the texture dataset of sampling box size 20, the first of the three configurations included pixel-neighboring of distance equal to one ($d = 1$). The best accuracy achieved was 55.1% only by the LSMD classifier, followed at 51.1% by the LDA classifier. The second configuration included all three available pixel distances ($d = 1,2,3$). The best accuracy rate achieved was 60.8% by both LDA and LSMD classifiers. Finally, a third configuration included all three available pixel distances ($d = 1,2,3$) plus all first-order statistics. The best accuracy achieved was 62.8% for both the LDA and the LSMD classifier. Table 4 summarizes the classification results for the 20-pixel box dataset.

For the texture dataset of sampling box size 50, the first of the three configurations included pixel-neighboring of distance equal to one ($d = 1$). The best accuracy achieved was 62.3% by both the LDA and the LSMD classifiers. The second configuration included all three available pixel distances ($d = 1,2,3$). The best accuracy rate achieved was 66.8% by both LDA and LSMD classifiers. Finally, a third configuration included all three available pixel distances ($d = 1,2,3$) plus all first-order statistics. The best accuracy achieved was 69.0% for both the LDA and the LSMD classifier. Table 4 summarizes the classification results for the 50-pixel box dataset.

3.2. Dataset fractal dimension

The calculation of the fractal dimension required the calculation of the correlation fractal dimension, i.e., the D_2 value, over each of the two texture datasets. Due to requirement for the best possible accuracy, when calculating the D_2 value, the analytical form of the pair-counting algorithm was chosen instead of the box-counting algorithm. Due to the increased storage and processing requirements of this method, instead of calculating the D_2 value over the entire dataset, it was averaged over multiple runs over smaller random subsets of 500 samples each. In all cases, the averaged D_2 value was confirmed to exhibit less than $\pm 3\%$ variation over all runs; thus it was used as a more realistic approximate when compared to the D_2 value, returned by the box-counting algorithm over the entire dataset.

Figs. 5 and 6 illustrate the PC(r) plots for the normalized texture datasets for sampling box sizes of 20 and 50 pixels, respectively. In both cases, no additional error weighting was required when calculating the fitness, i.e., the initial sigmoid function provided a fairly accurate estimation of the slope in the central part of the plot. A semi-rectangular ($q = 0.5$) Tukey window was required in the case of the qualitative features dataset (Table 1), in order to achieve optimal fitness for the sigmoid function in the central part of the curve.

Table 4 Classification results for LDA and LSMD classification against diagnosis, employing exhaustive combinations search through all the 1023 possible feature subsets of the 10 best MANOVA selections for the 20-pixel and 50-pixel box sizes

Texture dataset	20-Pixel box size		50-Pixel box size	
	LDA classifier (%)	LSMD classifier (%)	LDA classifier (%)	LSMD classifier (%)
Distances: $\{1\}$ first-order stats: NO	51.1	55.1	62.3	62.3
Distances: $\{1,2,3\}$ first-order stats: NO	60.8	60.8	66.8	66.8
Distances: $\{1,2,3\}$ first-order stats: YES	62.6	62.6	69.0	69.0

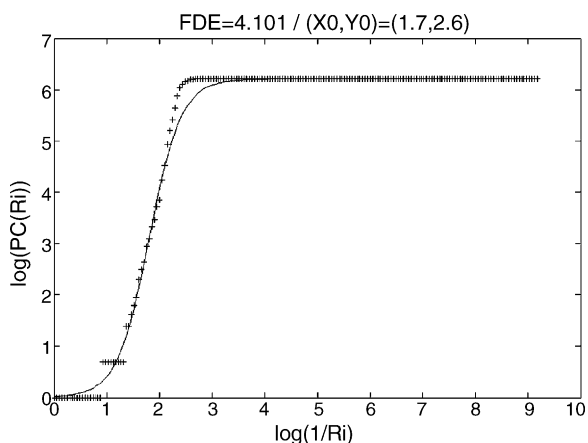


Figure 5 $PC(r)$ plot and sigmoid fitness function for the complete texture dataset ($\text{dim} = 120$), sampling box size of 20 pixels (1.270 mm). For X -axis, $1/r$ was used instead of r for correct (+) sign on the slope value.

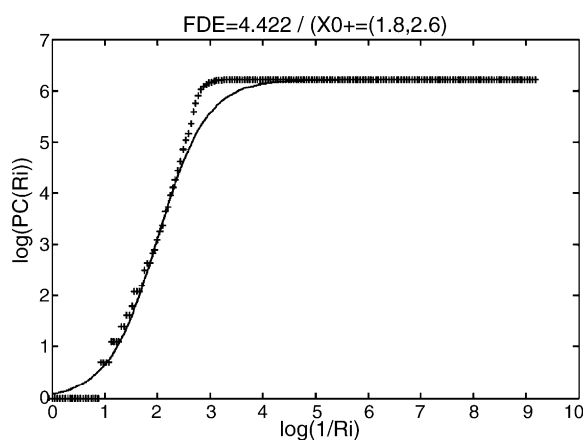


Figure 6 $PC(r)$ plot and sigmoid fitness function for the complete texture dataset ($\text{dim} = 120$), sampling box size of 50 pixels (3.175 mm). For X -axis, $1/r$ was used instead of r for correct (+) sign on the slope value.

Table 5 summarizes the results of D_2 value, i.e., the intrinsic fractal dimension, calculated over the two texture dataset, including all 120 features or using only the MANOVA selections for each dataset. The first column contains the D_2 values for the

complete dataset (120 features), while the second and the third columns contain the corresponding D_2 value when MANOVA is applied to select only the top-20 and top-10 statistically independent features, respectively. The calculated values are compared against the D_2 value of the original dataset of the qualitative features (Table 1), except the diagnosis itself. Although the exact values of these features are approximate quantifications provided by the expert physician, the corresponding D_2 value of this datasets can be used as a guideline. As the intrinsic fractal dimension essentially characterizes the descriptive power of the corresponding features [25], this D_2 value can be considered as the minimum inherent dimensionality of any textural or other dataset, required for describing the input samples with the same or higher level of quality, in terms of information content.

3.3. Classification performance

The third phase of the analysis included real classification setups against verified diagnosis, using linear and non-linear classification architectures. Specifically, the 10 best MANOVA feature selections, from the complete set of 120, were used as training sets for all classifier models, including LDA, LSMD, K-nn, RBF and MLPANNs, as well as SVMs. The choice of using only the top-10 MANOVA feature selections, instead of the top-20 or the entire 120-feature dataset, was based on the fact that the comparative fractal analysis of these datasets has proven the validity of using only a small subset of powerful features instead of all of them. Table 5 shows only small degradations in the descriptive power of such small subsets when compared to the complete feature set, which means that even the top-10 feature subset is enough to describe the full feature with great detail and complexity. Furthermore, the time requirements of using excessive input dimensionality in combination with leave-one-out cross-validation and sophisticated classifiers, like ANNs and SVMs, makes the training process impractical with little or no expected gain in the actual performance of the increased-size classifier.

Table 5 Correlation fractal dimension (D_2) value for the complete and sigmoid fitness function for the texture dataset of sampling box size of 20 pixels and 50 pixels, for all 120 features and for top-20 or top-10 MANOVA selected features

Fractal dim. D_2 value	Complete set ($\text{dim} = 120$)	MANOVA	
		Top-20	Top-10
Qualitative properties set	3.18	—	—
Textural features set, box size: 20 pixels	4.10	3.94	3.88
Textural features set, box size: 50 pixels	4.42	4.28	3.90

The two texture datasets for 20-pixel and 50-pixel box sizes were used separately in comparative training configurations, without restrictions over the pixel-neighboring distances ($d = 1, 2, 3$) in the MANOVA selections. All configurations included training patterns grouped according to their source image. In this way, despite the fact that each training pattern referred to only a local sample of the complete mass, it constituted one complete texture descriptor tagged with the appropriate “benign” or “malignant” diagnosis identifier.

In order to assess the true performance and generalization of each classifier model, the leave-one-out method was employed in all cases during training for both linear and non-linear classifiers. Specifically for ANN and SVM classifiers, instead of using the entire dataset, small subsets of 1000 random samples were used. The training and evaluation cycles were repeated multiple times, using a new random subset each time, and the final classification accuracy was calculated as the average over all runs, with standard deviation verified to be $\pm 1\%$ at most in all cases.

For the 20-pixel box dataset, the results of classification accuracy ranged from 62.6% to 80.4% according to the exact classifier selection. Both the LDA and LSMD classifiers achieved only the lowest performance of 62.6% even when using optimized combinations of features. Next, the MLP and RBF neural classifiers with optimized topology scored 74.4% and 71.3% correspondingly. The overall best accuracy was achieved by the SVM classifier at 80.4%, followed very closely by the optimized K-nn classifier ($k = 18$) at 80.3%.

For the 50-pixel dataset, the results of classification accuracy ranged from 69.0% to 83.9% according to the exact classifier selection. As in the first dataset, both LDA and LSMD achieved the lowest score equally at 69.0%. The RBF neural classifier achieved 72.8% and the MLP classifier outperformed it with better accuracy at 78.2%. Again, the overall best performance was achieved by the SVM classifier at 83.9%, followed closely by the optimized K-nn classifier ($k = 17$) at 83.6%.

It should be noted that, although ROC analysis of classifiers is common in medical applications, it was not employed in this study. The main property of ROC curves is their qualitative presentation of a classifier’s sensitivity and specificity levels for various decision thresholds [52]. However, their contribution become obscure in cases where clear quantitative and comparative results are needed for more than one classifier, or more than two output classes are involved [53]. In terms of a ROC curve, the best accuracy rate is represented by the point that exhibits the minimum distance from the ideal

Table 6 Best classification percentages of all classifiers for both texture datasets, using the top-10 MANOVA selections of features

Dataset classifier	20-Pixel box texture dataset (%)	50-Pixel box texture dataset (%)
LDA	62.6	69.0
LSMD	62.6	69.0
K-nn	80.3	83.6
NN/RBF	71.3	72.8
NN/MLP	74.4	78.2
C-SVC/RBF	80.4	83.9

Bold numbers indicate the overall best scores over each dataset.

classifier response, that is the point (0, 1) in the ROC space. Since this study is focused on investigating the discriminative power of textural features and their relative efficiency when used with various classifier architectures, the optimal performance was measured as the single value of their best accuracy rate that is calculated directly from the corresponding confusion matrix, instead of a ROC curve.

Table 6 summarizes the best scores of all classifiers for both texture datasets when the top-10 MANOVA selections of textural features were used.

4. Discussion

The extraction and analysis of localized textural features over the breast tissue is inherently related to the characterization of the tissue itself [12,13]. Thus, any qualitative characteristic of the clinical appearance of the tissue inside and around suspicious tumor areas can be investigated via texture analysis. This means that all of the clinical features included in Table 1, except the “external” data regarding the patient’s age, could be statistically analyzed and predicted on the basis of the texture of the digitized image over the corresponding tissue areas. However, the most important and practically useful configuration in clinical applications is the one that incorporates various data inputs and combines them in arbitrary schemes, in order to produce a valid and realistic assessment regarding the probable benignancy or malignancy of a suspicious breast tumor. As the topic of textural feature functions has been thoroughly investigated in other general or case-specific studies [33–36], this investigation was focused primarily on the appropriate use and optimization of the various textural feature functions when the target is the clinical assessment of suspicious tumors in mammograms.

4.1. Textural features extraction

There are three main issues related to the texture and the way it is processed in mammographic images: the sampling box size, the pixel-level processing scale and the statistical functions used to characterize the nature and type of the texture itself.

The results from the statistical significance analysis via MANOVA, as well as analytical classification runs using linear classifiers, have proven that the exact configuration of the textural feature functions is an issue of outmost importance. Tables 2 and 3 show that, in a total of 20 feature selections, only three of them refer to first-order statistics of the gray-level and only four of them refer to second-order or to run-length statistics using no box-interleaving, i.e., scaling factor. In Table 2, i.e., for sampling box size of 1.270 mm (20 pixels), there is only one first-order statistic (mean) and nine statistics of the co-occurrence or run-length matrices, from which only one refers to pixel-neighboring mode at distance $d = 1$. Similarly, in Table 3, i.e., for sampling box size of 3.175 mm (50 pixels), there are two first-order statistics (max, skewness) and eight statistics of the co-occurrence or run-length matrices, from which only three refer to pixel-neighboring mode at distance $d = 1$. Feature selections from both datasets clearly indicate that scale is a very important aspect of texture for the distinct characterization of benignancy or malignancy in local image processing. Furthermore, larger scales ($d = 2$ or $d = 3$) are preferred when a smaller sampling box is applied, while slightly smaller scales seem to be adequate if the sampling box is large enough to capture the same level of discriminating information when clinical diagnosis is concerned. Although there is no clear preference of one pixel-neighboring mode over the others, it is certain that the performance and quality of the textural feature functions in real diagnostic problems should be evaluated over multiple scales and configurations.

Results from Tables 2 and 3 are also suggestive with regard to the sampling box size and its effect on the information content captured. MANOVA selections indicate some preference over specific textural feature functions, like run-length non-uniformity, run-length short run emphasis and co-occurrence maximum correlation coefficient, but for various pixel distances or selections of mean versus the values' range. This means that *the exact selection of optimal feature sets can be estimated only in combination with specific sampling box sizes*. In other words, the size of the image area, over which the texture is analyzed, affects not only

the optimal scale by which the feature functions are applied, but the optimal selection of these feature functions as well. In terms of preliminary classification results, Table 4 clearly shows that in all cases the larger sampling box (50-pixel) is capable of capturing the diagnostic information in greater detail and quality than the smaller sampling box (20-pixel). This conclusion is verified by both linear classifiers, i.e., LDA and LSMD, as well as the slightly higher D_2 value for the fractal dimension of the corresponding datasets (Table 5).

In most cases, the mean value of each textural feature was preferred instead of its values' range, regardless of the exact feature function. The three first-order statistics included in Tables 2 and 3 are all directly (mean, max) or indirectly (skewness) related to the inherent bias of the tumor's histogram towards black or white values, i.e., towards more fatty or dense tissue, correspondingly. Regarding the second-order statistics, there is no clear preference between co-occurrence and run-length based feature functions, regardless of sampling box size and pixel-neighboring modes.

4.2. Dataset fractal analysis over feature selections

The results presented in Table 5 clearly indicate that the complete texture datasets, containing 120 features each, exhibit clearly higher D_2 values against the eight qualitative features in all cases. In other words, all texture datasets, even the ones reduced to top-20 features by applying MANOVA for optimal feature selection, employ at least the same or higher descriptive power as the qualitative dataset does.

In all cases, the differences between the D_2 value of the complete datasets (dim = 120) against the MANOVA reduced subsets (dim = 20 or 10) is less than -9%, for the sampling box sizes of both 20 and 50 pixels. This essentially means that, *even when using only a small subset of optimal textural features instead of the complete set, the intrinsic descriptive power of the resulting datasets remains high*. This requirement is crucial when designing minimal classifier models for real diagnostic applications. As a result, the MANOVA analysis for the selection of the 10 or 20 highly uncorrelated textural features was considered safe and consistent, for constructing training datasets of minimal dimensionality and similar descriptive power, for all the classifier test models employed in this study.

On the other hand, the corresponding D_2 value when using only the best 20 or 10 MANOVA selections of features should not be characterized as conclusive for describing the complete texture datasets of dimension 120. Although the D_2 values indicate an

intrinsic dimensionality lower than 5, the descriptive power of the reduced datasets, in terms of intrinsic fractal dimension, can be observed even when using the best 20 or 10 MANOVA selections, with D_2 reducing down to -4% and -9% , respectively. Furthermore, non-fractal datasets produce $PC(r)$ plots of limited or no true linear sections for the calculation of slope, i.e., of the D_2 value, especially when the distribution of the samples is sparse or clustered [43]. As a result, individual feature inclusions or exclusions produce minimal fluctuations of the slope in the $PC(r)$ plot, sometimes smaller than the fitness error of the sigmoid itself. Therefore, the fractal methods for the complexity analysis of a dataset is not always a safe and conclusive means for providing exact and optimal combinations of feature selections for classification purposes.

4.3. Classifier models evaluation

Conclusive classification results proved the value of non-linear architectures versus the linear models. Results from Table 4 demonstrate the almost identical performance of LDA and LSMD classifiers, in all cases except one (for $d=1$). In both cases, the relatively low accuracy rates over the two texture datasets, i.e., 62.6% for the 20-pixel box and 69.0% for the 50-pixel box sizes, proved that the inherent complexity of the problem is clearly non-linear.

Concerning neural networks, comparison between RBF and MLP architectures proved that RBF networks resulted in somewhat lower overall accuracy for both texture datasets. Differences in success rates ranged between 4% and 7% , marginally favoring the choice of MLP over networks of similar topologies. The largest network topologies that were used included neural layers of sizes 10-6-2 (input-hidden-output), i.e., a total of 80 wt. parameters and a samples-to-weights ratio of 12.5:1. This restriction on topology sizes was enforced in order to assure a relatively high generalization level and realistic classification results [31]. The overall best for neural classifiers of maximal topology was achieved by MLP at 78.2% , using the 50-pixel box texture dataset.

A very interesting conclusion drawn from Table 6 is the fact that the K-nearest neighbor classifier achieved only marginally lower accuracy rates compared to the SVM classifier. For the 20-pixel box texture dataset this difference was only -0.1% with SVM at the overall best of 80.4% , while for the 50-pixel box texture dataset this difference was only -0.3% , with SVM at the overall best of 83.9% . It should be noted that in both cases the K-nn classifier employed a relatively large K value, specifically $K=18$ and $K=17$, respectively. This essentially means that the K-nn classifier required a relatively

large sample of local “neighbors” in order to produce an accurate class prediction. The K-nn approaches the statistically optimal prediction because for large values of K the majority of the points in the local “neighborhood” will belong to the class corresponding to the *maximum conditional probability* [31,54]. In other words, as the statistical sample of the local “neighborhood” increases in size, the classification selections and overall performance of the K-nn classifier tends to approximate the Bayesian model for the same problem. The SVM model embodies a similar optimality criterion, but in a structural way. Specifically, the SVM is trained in a way that the final classifier exhibits the property of incorporating the *maximum margin* possible for the decision boundary between the two classes [32,47,55]. In other words, the trained SVM classifier employs the best separation criterion when predicting the correct class for an unknown sample, i.e., inherently the optimal *generalization* criterion [31,56]. As a result of having unbiased training datasets, the accuracy rates of optimal K-nn and optimal SVM classifiers differed only slightly, favoring the SVMs in all cases.

5. Conclusion

Texture analysis is one of the most valuable and promising areas in breast tissue analysis and characterization. Extensive clinical studies and interaction with human expert physicians have established a wide range of mammographic properties, which are considered of utmost importance in the estimation of the clinical status and the formulation of a robust diagnostic model. Most of these diagnostic features are related to textural properties of the mammographic image, thus they can be investigated in a quantitative and systematic way via automated texture extraction and analysis, directly from the digitized mammographic image.

Extensive studies of these textural features in various scales and configurations are necessary in order to select a useful set of texture descriptors, specialized for the specific task of breast mass tissue characterization. The sampling box size and pixel-neighboring scheme proved to be very important factors in the optimization and final selection of the various textural feature functions. Analysis of the intrinsic dimensionality of the resulting texture datasets have established their robustness in terms of descriptive power, similar to the qualitative features that were provided by the expert physician. Consequently, further studies have to be conducted towards the formulation and application of scalable image texture descriptors, like 2D wavelet decomposition structures [57] and fractal analysis of texture

[42,58–60], especially focused on the diagnostic problem of mammographic mass characterization.

The application of robust classifier models proved to be of outmost importance as well. Classification results over linear models, employing exhaustive feature combination optimization, have provided some indications regarding the most appropriate textural feature functions and their configurations for local processing, for classification problems against the verified diagnosis of the complete tumor. As the problem becomes too complex for simple linear systems, more efficient structures are necessary in order to exploit the complete range of the discriminating power of the available texture datasets. MLP neural classifiers outperformed all other linear and ANN architectures, while K-nn and especially SVM classifiers achieved the overall best accuracy rates.

As the texture involves only a portion of the complete information content of mammographic images, the prospect of using texture in conjunction with other methodologies, like structural or morphological mass analysis, into a combined diagnostic tool, is very promising.

Appendix A. Textural features

A.1. First-order gray-level statistics

Denoting by $l(x,y)$ the image sub-region pixel matrix, the formulae used for the standard statistics of the gray-level are as follows:

1. Min:

$$l_{\min} = \min_{XY} \{l(x,y)\} \quad (\text{A.1})$$

2. Max:

$$l_{\max} = \max_{XY} \{l(x,y)\} \quad (\text{A.2})$$

3. Mean:

$$\mu = \frac{1}{XY} \sum_{x=1}^X \sum_{y=1}^Y l(x,y) \quad (\text{A.3})$$

4. Variance:

$$\sigma^2 = \frac{1}{(XY-1)} \sum_{x=1}^X \sum_{y=1}^Y [l(x,y) - \mu]^2 \quad (\text{A.4})$$

5. Skewness:

$$\frac{1}{XY} \sum_{x=1}^X \sum_{y=1}^Y \left[\frac{l(x,y) - \mu}{\sigma} \right]^3 \quad (\text{A.5})$$

6. Kurtosis:

$$\left\{ \frac{1}{XY} \sum_{x=1}^X \sum_{y=1}^Y \left[\frac{l(x,y) - \mu}{\sigma} \right]^4 \right\} - 3 \quad (\text{A.6})$$

A.2. Second-order gray-level statistics

Denoting by $p(i,j)$ the normalized co-occurrence matrix, by N_g the number of discrete gray levels of the images, by $p_x(i)$ and $p_y(j)$ the row and column marginal probabilities, respectively, obtained by summing the columns or rows of $p(i,j)$:

$$\text{i.e., } p_x(i) = \sum_{j=1}^{N_g} p(i,j) \quad \text{and} \quad (\text{A.7})$$

$$p_y(j) = \sum_{i=1}^{N_g} p(i,j)$$

$$p_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{\substack{j=1 \\ i+j=k}}^{N_g} p(i,j) \quad k = 2, 3, \dots, 2N_g \quad (\text{A.8})$$

$$p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{\substack{j=1 \\ |i-j|=k}}^{N_g} p(i,j), \quad k = 0, 1, \dots, N_g \quad (\text{A.9})$$

the formulae used for the metrics of the SDGM are as follows:

1. Angular second momentum (ASM):

$$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \{p(i,j)\}^2 \quad (\text{A.10})$$

2. Contrast:

$$\sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{\substack{i=1 \\ |i-j|=n}}^{N_g} \sum_{j=1}^{N_g} p(i,j) \right\} \quad (\text{A.11})$$

3. Correlation:

$$\frac{\left[\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (ij) p(i,j) \right] - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (\text{A.12})$$

where

$$\mu_x = \sum_{i=1}^{N_g} \left[i \sum_{j=1}^{N_g} p(i,j) \right], \quad (\text{A.13})$$

$$\mu_y = \sum_{j=1}^{N_g} \left[j \sum_{i=1}^{N_g} p(i,j) \right], \quad (\text{A.14})$$

$$\sigma_x = \sum_{i=1}^{N_g} \left[(i - \mu_x)^2 j \sum_{j=1}^{N_g} p(i,j) \right], \quad (\text{A.15})$$

$$\sigma_y = \sum_{j=1}^{N_g} \left[(j - \mu_y)^2 \sum_{i=1}^{N_g} p(i,j) \right] \quad (\text{A.16})$$

are the mean values and standard deviations of p_x and p_y , respectively.

4. Sum of squares–variances:

$$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} [(i - \mu)^2 p(i, j)] \quad (\text{A.17})$$

5. Inverse difference moment:

$$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \left[\frac{1}{1 + (i - j)^2} p(i, j) \right] \quad (\text{A.18})$$

6. Sum average:

$$\sum_{i=2}^{2N_g} [i p_{x+y}(i)] \quad (\text{A.19})$$

7. Sum variance:

$$\sum_{i=2}^{2N_g} [(i - \text{sum average})^2 p_{x+y}(i)] \quad (\text{A.20})$$

8. Sum entropy:

$$-\sum_{i=2}^{2N_g} [p_{x+y}(i) \log[p_{x+y}(i)]] \quad (\text{A.21})$$

9. Entropy:

$$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} [p(i, j) \log(p(i, j))] \quad (\text{A.22})$$

10. Difference variance: i.e., variance of

$$p_{x-y} = \sum_{i=0}^{N_g-1} [(i - f')^2 p_{x-y}(i)] \quad (\text{A.23})$$

where

$$f' = \sum_{i=0}^{N_g-1} [i p_{x-y}(i)] \quad (\text{A.24})$$

11. Difference entropy:

$$-\sum_{i=0}^{N_g-1} [p_{x-y}(i) \log[p_{x-y}(i)]] \quad (\text{A.25})$$

12. Information measures of correlation:

$$\frac{\text{HXY} - \text{HXY1}}{\max\{\text{HX}, \text{HY}\}} \quad (\text{A.26})$$

and

$$[1 - \exp[-2.0(\text{HXY2} - \text{HXY})]]^{1/2} \quad (\text{A.27})$$

where HX and HY is the entropy of p^x and p^y , respectively and

$$\text{HXY} = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} [p(i, j) \log(p(i, j))] \quad (\text{A.28})$$

$$\begin{aligned} \text{HXY1} = & -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} [p \\ & \times (i, j) \log(p_x(i) p_y(j))] \end{aligned} \quad (\text{A.29})$$

$$\text{HXY2} = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} [p_x(i) p_y(j) \log(p_x(i) p_y(j))] \quad (\text{A.30})$$

13. Maximal correlation coefficient: (second greatest eigenvalue of Q)^{1/2}, where

$$Q(i, j) = \sum_k \frac{p(i, k) p(j, k)}{p_x(i) p_y(k)} \quad (\text{A.31})$$

A.3. Run-length gray-level statistics

Denoting by P the total number of pixels of an image, by $p(i, j)$ the (i, j) -th element of the run-length matrix for a specific angle θ and a specific distance d (i.e., $p_{\theta,d}(i, j)$) and by N_r the number of different run lengths that occur, the formulae used are as follows:

1. Short run emphasis:

$$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i, j) / j^2}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i, j)} \quad (\text{A.32})$$

2. Long runs emphasis:

$$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} j^2 p(i, j)}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i, j)} \quad (\text{A.33})$$

3. Gray-level non-uniformity:

$$\frac{\sum_{i=1}^{N_g} \left[\sum_{j=1}^{N_r} p(i, j) \right]^2}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i, j)} \quad (\text{A.34})$$

4. Run length non-uniformity:

$$\frac{\sum_{j=1}^{N_r} \left[\sum_{i=1}^{N_g} p(i, j) \right]^2}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i, j)} \quad (\text{A.35})$$

5. Run percentage:

$$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i, j)}{P} \quad (\text{A.36})$$

Appendix B. Tukey window

The Tukey windows [44,45] are cosine-tapered functions. They are parametric against q , where q specifies the exact form of the window, ranging from completely rectangular ($q = 0$) to completely triangular or Hanning ($q = 1$). In this study, all fitness

calculations employed Tukey windows as error weighting factors, using parameters q between 0.5 and 1.0 for optimal results.

The equation for computing the coefficients $w[k]$ of a discrete Tukey window is as follows:

$$w[k] = \begin{cases} \frac{1}{2} \left(1 + \cos \left(\frac{2\pi(k-1)}{q(N-1)} - \pi \right) \right), & 1 \leq k < \frac{q}{2}(N-1) \\ 1, & \frac{q}{2}(N-1) \leq k \leq N - \frac{q}{2}(N-1) \\ \frac{1}{2} \left(1 + \cos \left(\frac{2\pi}{q} - \frac{2\pi(k-1)}{q(N-1)} - \pi \right) \right), & N - \frac{q}{2}(N-1) < k \leq N \end{cases} \quad (\text{A.37})$$

References

- [1] Newstead GM, Baute PB, Toth HK. Invasive lobular and ductal carcinoma: Mammographic findings and stage at diagnosis. *Radiology* 1992;184:623–7.
- [2] Meyer JE, Amin E, Lindfors KK, Lipman JC, Stomper PC, Genest D. Medullary carcinoma of the breast: Mammographic and US appearance. *Radiology* 1989;170:79–82.
- [3] Sickles EA. Breast masses: mammographic evaluation. *Radiology* 1989;173:297–303.
- [4] Egan RL. Breast imaging: diagnosis and morphology of breast diseases Philadelphia: Saunders; 1988.
- [5] Robbins SL, Angell M, Kumar V. Basic pathology Philadelphia: Saunders; 1981.
- [6] Bocchi L, Coppini G, De Dominicis R, Valli G. Tissue characterization from X-ray images. *Med Eng Phys* 1997;19(4):336–42.
- [7] Homer MJ. Imaging features and management of characteristically benign and probably benign lesions. *Radiol Clin N Am* 1987;25(5):939–51.
- [8] Homer MJ. Breast imaging: Pitfalls, controversies and some practical thoughts. *Radiol Clin N Am* 1985;23(3):459–72.
- [9] Tomita F, Tsuji S. Computer analysis of visual textures Massachusetts, USA: Kluwer Academic Publishers; 1990.
- [10] D’Orsi CJ, Getty DJ, Swets JA, Pickett RM, Seltzer SE, McNeil BJ. Reading and decision aids for improved accuracy and standardization of mammographic diagnosis. *Radiology* 1992;184:619–22.
- [11] Wu Y, Giger ML, Doi K, Vyborny CJ, Schmidt RA, Metz CE. Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer. *Radiology* 1993;187:81–7.
- [12] Petersen EM, Ridder D, Handels H. Image processing with neural networks – a review. *Pat Rec* 2002;35:2279–301.
- [13] Li S, Kwok JT, Zhu H, Wang Y. Texture classification using the support vector machines. *Pat Rec* 2003;36:2883–93.
- [14] Olson SL, Fam BW, Winter PF, Scholz FJ, Lee AK, Gordon SE. Breast calcifications: analysis of imaging properties. *Radiology* 1988;169:329–32.
- [15] Lanyi M. Diagnosis and differential diagnosis of breast calcifications Berlin: Springer-Verlag; 1988.
- [16] Dhawan AP, Chitre Y, Kaiser-Bonasso C. Analysis of mammographic microcalcifications using gray-level image structure features. *IEEE Trans Med Im* 1996;15(3):246–59.
- [17] Kim JK, Park HW. Statistical textural features for detection of microcalcifications in digitized mammograms. *IEEE Trans Med Im* 1999;18(3):231–8.
- [18] Strickland RN, Hahn HI. Wavelet transforms for detecting microcalcifications in mammograms. *IEEE Trans Med Im* 1996;15(2):218–29.
- [19] Lisboa PJG. A review of evidence of health benefit from artificial neural networks. *Neural Networks* 2002;15:11–39.
- [20] Arivazhagan S, Ganesan L. Texture classification using wavelet transform. *Pat Rec Lett* 2003;24:1513–21.
- [21] Li H, Wang Y, Liu KJR, Lo S-CB, Freedman MT. Computerized radiographic mass detection. Part I. Lesion site selection by morphological enhancement and contextual segmentation. *IEEE Trans Med Im* 2001;20(4):289–301.
- [22] Sahiner B, Petrick N, Chan HP, Hadjiiski LM, Paramagul C, Helvie MA, Gurcan MN. Computer-aided characterization of mammographic masses: accuracy of mass segmentation and its effects on characterization. *IEEE Trans Med Im* 2001;20(12):1275–84.
- [23] Hatanaka Y, Hara T, Fujita H, Kasai S, Endo T, Iwase T. Development of an automated method for detecting mammographic masses with a partial loss of region. *IEEE Trans Med Im* 2001;20(12):1209–14.
- [24] Cheng HD, Muiy Cui. Mass lesion detection with fuzzy neural network. *Pat Rec* 2004;37:1189–900.
- [25] Traina Jr C, Traina AJM, Wu L, Faloutsos C. Fast feature selection using fractal dimension. In: XV Brazilian Database Symposium, vol. 1. 2000. p. 158–71.
- [26] Mavroforakis M, Georgiou H, Dimitropoulos N, Cavouras D, Theodoridis S. Significance analysis of qualitative mammographic features, using linear classifiers, neural networks and support vector machines. *Eur J Radiol* 2005;54:80–9.
- [27] Cooley WW, Lohnes PR. Multivariate data analysis New York: John Wiley and Sons; 1971.
- [28] Leeden R, Vrijburg K, Leeuw J. A review of two different approaches for the analysis of growth data using longitudinal mixed linear models: comparing hierarchical linear regression (ML3, HLM) and repeated measures designs with structured covariance matrices (BMDP5V). *The Stat Software Newslett SSNinCSDA* 1996;20:583–605.
- [29] Ackerman LV, Mucciardi AN, Gose EE, Alcorn FS. Classification of benign and malignant breast tumors on the basis of 36 radiographic properties. *Cancer* 1973;31:342–52.
- [30] Suckling J, Parker J, Dance DR, Astley S, Hutt I, Daggis CRM, Ricketts I, Stamatakis E, Cerneaz N, Kok SL, Taylor P, Betal D, Savage J. The mammographic image analysis society digital mammogram database, exerpta medica. In: International Congress Series, vol. 1069. 1994. p. 375–8.
- [31] Theodoridis S, Koutroumbas K. Pattern recognition, 2nd ed., San Diego, USA: Academic Press; 2003.
- [32] Christianini N, Shawe-Taylor J. An introduction to support vector machines UK: Cambridge University Press; 2000.
- [33] Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Sys Man Cyb SMC* 1973;3(3):610–21.
- [34] Haralick RM. Statistical and structural approaches to texture. *Proc IEEE* 1979;67(5):786–804.
- [35] Galloway M. Texture analysis using gray level run lengths. *Comp Graph Im Proc* 1975;4:172–9.
- [36] Gonzalez RC, Woods RE. Digital image processing New York: Addison-Wesley; 1992.

- [37] Stone M. Cross-validators choice and assessment of statistical predictions. *J R Stat Soc B* 1974;36(1):111–47.
- [38] Moore AW, Lee MS. Efficient algorithm for minimizing cross validation error. In: *Proceedings of the 11th International Conference M.L.*. San Francisco: Morgan Kaufmann; 1994 .
- [39] Lachenbruch PA, Mickey RM. Estimation of error rates in discriminant analysis. *Technometrics* 1968;10:1–11.
- [40] Chen CC, Daponte JS, Fox MD. *IEEE Trans Med Im* 1989;8(2): 133–42.
- [41] Abrahao B, Barbosa L. *Characterizing datasets using fractal methods* Belo Horizonte, MG Brazil: Department of Computer Science, Universidade Federal de Minas Gerais; 2003 .
- [42] Pierre S, Rivest JF. On the validity of fractal dimension measurements in image analysis. *J Vis Com Im Proc* 1996; 7(3):217–29.
- [43] Massopust PR. *Fractal functions fractal surfaces and wavelets* San Diego: Academic Press; 1994.
- [44] Harris FJ. On the use of windows for harmonic analysis with the Discrete Fourier Transform. *Proc IEEE* 1978;66:66–7.
- [45] MathWorks Inc.. *Matlab 7.0 documentation: signal processing toolbox* MathWorks, Mass; 2004.
- [46] Devroye L, Györfi L, Lugosi G. *A probabilistic theory of pattern recognition* New York: Springer-Verlag Inc.; 1996.
- [47] Theodoridis S. Pattern recognition (lemma). In: Bigdoli H, editor. *Encyclopedia of Information Systems*, vol. 3. Chestnut Hill, MA: Academic Press; 2003. p. 459–79.
- [48] Poggio T, Girosi F. Networks for approximation and learning. *Proc IEEE* 1990;78:1481–97.
- [49] Haykin S. *Neural networks: a comprehensive foundation*, 2nd ed., New Jersey: Prentice Hall; 1999.
- [50] Bennett KP, Campbell C. Support vector machines: hype or hallelujah? *ACM SIGKDD Explor* 2000;2(2):1–13.
- [51] Reed R. Pruning algorithms – a survey. *IEEE Trans N N* 1998; 4(5):740–7.
- [52] Swets JA, Pickett RM. *Evaluation of diagnostic systems: methods from signal detection theory* New York: Academic Press; 1982.
- [53] Fawcett T. *ROC Graphs: notes and practical considerations for data mining researchers*. Intelligent Enterprise Technologies Laboratory, HP Laboratories Palo Alto, HPL-2003-4; 2003.
- [54] Pernkopf F. Bayesian network classifiers versus selective K-nn classifier. *Pat Rec* 2005;38:1–10.
- [55] Vapnik VN. *Statistical learning theory* New York: John Wiley and Sons; 1998.
- [56] Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Disc* 1998;2(2):1–47.
- [57] Vetterli M, Herley C. Wavelets and filter banks: theory and design. *IEEE Trans Sig Proc* 1992;40(9):2207–32.
- [58] Encarnacao JL, Peitgen H-O, Sakas G, Englert G. *Fractal geometry and computer graphics* Berlin: Springer-Verlag; 1992.
- [59] Li H, Liu KJR, Lo SCB. Fractal modeling and segmentation for the enhancement of microcalcifications in digital mammograms. *IEEE Trans Med Im* 1997;16(6):785–98.
- [60] Caldwell CB, Stapleton SJ, Holdsworth DW, Jong RA, Weiser WJ, Cooke G, Yaffe MJ. Characterization of mammographic parenchymal pattern by fractal dimension. *Phys Med Biol* 1990;35:235–47.